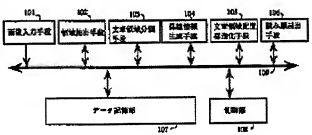
DOCUMENT PICTURE PROCESSING METHOD

Publication number: JP7192084
Publication date: 1995-07-28
Inventor: SAITO TAKASHI
Applicant: RICOH KK
Classification:
- international: ***G06K9/20; G06K9/62;*** *G06K9/20;* **G06K9/20; G06K9/62;** G06K9/20; (IPC1-7): G06K9/20; G06K9/62
- European:
Application number: JP19930327015 19931224
Priority number(s): JP19930327015 19931224; JP19930110397 19930512; JP19930288960 19931118

Abstract of **JP7192084**

PURPOSE:To highly accurately structure the documtent areas of vertically written and horizontally written documents and to extract a correct reading order. CONSTITUTION:An area extraction means 102 extracts the areas of character areas and drawing areas, etc., from binary pictures, a sentence area classification means 103 performs classification into drawing titles, titles, headers, footers and the other text areas and a ruled line information generation means 104 generates the imaginary ruled lines of an extracted ruled line area and a white area and the imaginary ruled lines of the end part of the drawing area, etc. A sentence area arrangement structuring means 105 structures the arrangement of the text area and expresses it by a tree graph and a reading order extraction means 106 decides the reading order from the graph expression.



Back to JP719

Family list
4 family members for: **JP7192084**
Derived from 3 applications

1 **DOCUMENT PICTURE PROCESSING METHOD**
Inventor: SAITO TAKASHI
Applicant: RICOH KK
EC:
IPC: ***G06K9/20***; ***G06K9/62***; *G06K9/20* (+5)
Publication info: **JP3302147B2 B2** - 2002-07-15
JP7192084 A - 1995-07-28

2 **Document image processing method and system having function of determining body text region reading order**
Inventor: SAITOH TAKASHI (JP)
Applicant: RICOH KK (JP)
EC: G06K9/20L; G06K9/20R
IPC: ***G06K9/20***; **G06K9/20**; (IPC1-7): G06K9/34
Publication info: **US5774580 A** - 1998-06-30

3 **Document image processing method and system having function of determining body text region reading order**
Inventor: SAITOH TAKASHI (JP)
Applicant: RICOH KK (JP)
EC: G06K9/20L3
IPC: ***G06K9/20***; **G06K9/20**; (IPC1-7): G06K9/34
Publication info: **US5907631 A** - 1999-05-25

Data supplied from the ***esp@cenet*** database - Worldwide

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報(A)　(11)特許出願公開番号

特開平7－192084

(43)公開日　平成7年(1995)7月28日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｋ	9/20	３４０	Ｌ		
	9/62		Ｚ 9289－5Ｌ		

審査請求　未請求　請求項の数8　ＯＬ　（全 13 頁）

(21)出願番号　特願平5－327015

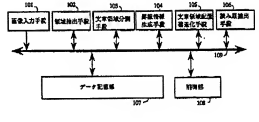
(22)出願日　平成5年(1993)12月24日

(31)優先権主張番号　特願平5－110397
(32)優先日　平5(1993)5月12日
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平5－288960
(32)優先日　平5(1993)11月18日
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000006747
株式会社リコー
東京都大田区中馬込１丁目３番６号
(72)発明者　齋藤　高志
東京都大田区中馬込１丁目３番６号　株式会社リコー内
(74)代理人　弁理士　鈴木　誠　（外１名）

(54)【発明の名称】　文書画像処理方法

(57)【要約】
【目的】　縦書き、横書き文書の文章領域を高精度に構造化して、正しい読み順を抽出する。
【構成】　領域抽出手段１０２は、２値画像から文字領域、図領域などの領域を抽出する。文章領域分別手段１０３は、図題、表題、ヘッダ、フッタと、それ以外の本文領域に分類する。罫線情報生成手段１０４は、抽出された罫線領域や、白領域の架空罫線、図領域の端部の架空罫線などを生成する。文章領域配置構造化手段１０５は、本文領域の配置を構造化し、木グラフで表現し、読み順抽出手段１０６は、このグラフ表現から読み順を決定する。



1

【特許請求の範囲】

【請求項1】　入力された文書画像から文章領域を抽出し、該抽出された文章領域を、本文領域とそれ以外の領域に分け、該本文領域の配置構造を求め、該配置構造を基に前記本文領域の読み順を抽出することを特徴とする文書画像処理方法。

【請求項2】　前記文書画像が縦書きまたは横書き書式であるとき、該縦書き横書き書式に共通な本文領域の配置構造を求め、該配置構造を基に該本文領域の読み順を抽出することを特徴とする請求項1記載の文書画像処理方法。

【請求項3】　前記配置構造を木グラフで表現し、前記各本文領域を木グラフの各ノードに割り当てることを特徴とする請求項1記載の文書画像処理方法。

【請求項4】　前記文章領域を囲み枠内の本文領域と囲み枠外の本文領域に分け、該囲み枠内の本文領域は、囲み枠外の本文領域とは別に読み順を抽出することを特徴とする請求項1記載の文書画像処理方法。

【請求項5】　前記抽出された読み順を評価し、該評価が偽と判定されたとき該読み順の再設定を行うことを特徴とする請求項1記載の文書画像処理方法。

【請求項6】　前記読み順の評価は、前記各本文領域に基準点を設けて、該基準点間を読み順に従って線分で結んだとき、線分に交わりが生じた場合に偽と判定することを特徴とする請求項5記載の文書画像処理方法。

【請求項7】　囲み枠内の本文領域は、それぞれの枠内の領域について読み順の評価と再設定を行うことを特徴とする請求項5記載の文書画像処理方法。

【請求項8】　前記読み順に従って、前記各本文領域に対して文字認識を行うことを特徴とする請求項1記載の文書画像処理方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、縦書きまたは横書き何れか一方の書式で書かれた文書画像から抽出された文章領域を構造化して、読み順を得るようにした文書画像処理方法に関する。

【0002】

【従来の技術】ＯＣＲ装置の前処理あるいは文書データベース構築の前処理においては、文書画像から抽出した文章領域の読み順を求める必要がある。従来、文章領域の読み順を求める方法としては、例えば、文章領域を点座標などによって並べて初期状態を得て、この状態から隣合う並びの領域同志について、判定手段を用いて入れ替えを行っていくことによって、最終的に読み順に領域が並び合うようにし、また読み取り順序の初期状態を得る際に、罫線などの非文字領域を追加することにより、読み取り順序の指定を容易した文書読み取り装置が提案されている（特開平3－269689号公報を参照）。

【0003】他の方法として、文書の領域分割を行った

2

後に、同一段組にある文章領域をノードとする木グラフを作成し、このグラフから論理的構造を得て、文章画像中の記事を読み順に従って抽出する文章画像処理装置が提案されている（特開平1－183784号公報を参照）。

【0004】

【発明が解決しようとする課題】上記した第1の技術は、文章領域およびそれ以外の領域を座標などを用いて並べて初期状態を得て、この状態から並び合う領域同志を比較し、読み順が逆になっていると判断したとき、並び替えるものである。この方法では、初期状態に依存し、しかも隣合う領域同志を比較していのので、タイトル部位のように本文と離れている場合には初期状態で隣合わない上に、２つの領域だけを比較した場合に、どちらが読み順が先になるかが局所的には定まらないときには、最終的にタイトル部位と本文との比較が行われない状態に陥る。また、非文字領域について、文字領域と同様に扱う場合のみしか考慮していないので、文字列方向と垂直方向の罫線情報や図があった場合の影響などが考慮されていない。

【0005】また、上記した第２の技術は、前提として段組を考慮しているので、明確な段組がない場合や、頁の上部が２段、下部が３段の如く、変則的な段組がなされていた場合に、「同一段組の文字領域を一つのノードに相当させる」ということが難しい。この方式もまた、文字方向と垂直方向の罫線があった場合や図表の影響が考慮されていない。

【0006】また、第１の技術は、行方向縦を前提として、読み順は隣合うブロックの場合に右から左へ進み、第２の技術では文字方向横を前提として上下に並ぶ領域を上から下へと並ぶようにノード内の順番を決めている。つまり、両者共にどちらか一方の行方向のみに対処している。また、何れも図表、表題やヘッダ、フッタといった本文領域とは異なる性質を持った文字領域の影響を全く考慮したものではない。

【0007】本発明の目的は、縦書きまたは横書き文書の文章領域を高精度に構造化して、正しい読み順を抽出する文書画像処理方法を提供することにある。

【0008】本発明の他の目的は、本文領域の構造化に失敗した場合に、読み順の再設定を行い、利用者による修正負担を少なくした文書画像処理方法を提供することにある。

【0009】本発明の更に他の目的は、正しく設定された読み順に従って文書情報を抽出して利用する文書画像処理方法を提供することにある。

【0010】

【課題を解決するための手段】前記第１の目的を達成するために、請求項１記載の発明では、入力された文書画像から文章領域を抽出し、該抽出された文章領域を、本文領域とそれ以外の領域に分け、該本文領域の配置構造

を求め、該配置構造を基に前記本文領域の読み順を抽出することを特徴としている。

【0011】請求項2記載の発明では、前記文書画像が縦書きまたは横書き書式であるとき、該縦書き横書き書式に共通な本文領域の配置構造を求め、該配置構造を基に該本文領域の読み順を抽出することを特徴としている。

【0012】請求項3記載の発明では、前記配置構造をグラフで表現し、前記各本文領域を木グラフの各ノードに割り当てることを特徴としている。

【0013】請求項4記載の発明では、前記文章領域を囲み枠内の本文領域と囲み枠外の本文領域に分け、該囲み枠内の本文領域は、囲み枠外の本文領域とは別に読み順を抽出することを特徴としている。

【0014】前記第2の目的を達成するために、請求項5記載の発明では、前記抽出された読み順を評価し、該評価が偽と判定されたとき該読み順の再設定を行うことを特徴としている。

【0015】請求項6記載の発明では、前記読み順の評価は、前記各本文領域に基準点を設けて、該基準点間を読み順に従って線分で結んだとき、線分に交わりが生じた場合に偽と判定することを特徴としている。

【0016】請求項7記載の発明では、囲み枠内の本文領域は、それぞれの枠内の領域について読み順の評価と再設定を行うことを特徴としている。

【0017】前記第3の目的を達成するために、請求項8記載の発明では、前記読み順に従って、前記各本文領域に対して文字認識を特徴としている。

【0018】

【作用】実施例1では、まず文書のイメージより抽出された文章（文字）領域を本文領域とそれ以外の領域とに分別する。ここで、本文領域とは読み順が設定されるべき文章領域のことで、それ以外の文章領域とは図題、表題、ヘッダ、フッタ等の本文領域とは異なった性質を持つ領域のことである。

【0019】そして、本文領域の配置構造を木グラフで表現する。この木グラフの構築にあたって、文字行方向に相対する座標系を採用するとともに文章領域を各ノードに対応させ、また、各ノードの勢力範囲というものを垂直罫線等を利用しながら規定し、この勢力範囲に従ってノード間の親子関係の探索を行なう。さらに、囲み枠内の文章領域については、囲み枠毎に同様の木グラフを求め、これを全体の木グラフに結合する。このようにして構築した本文領域の配置構造の木グラフ上で、本文領域の先行順探索を行なうことにより、本文領域の読み順を設定する。

【0020】文章領域を本文領域とそれ以外の領域とに分別することによって、図題等の本文領域以外の文章領域の影響を取り除いた本文領域の木グラフ表現を得られる。したがって、この木グラフに従って、本文領域とは異なる読み順を持つ文章領域に影響させることなく、本文領域の正しい読み順を設定できる。

【0021】文字行方向に相対する座標系を採用することにより、文字行方向が縦でも横でも木グラフを共通に扱うことができるようになるため、縦書き書式の文書も横書き書式も同様に処理可能になる。

【0022】本文領域を木グラフのノードに対応させることによって、段組に依存しない、本文領域の大局的な構造化が可能である。しかも、各ノードの勢力範囲を規定し、これを利用してノードの親子関係探索を行なうため、単に上下に並んでいるか否かといった単純な処理方法では対処不可能であった、タイトル部位や図等の存在に対処して本文領域の正確な構造化が可能であり、したがって、より正確な本文領域の読み順設定が可能である。

【0023】実施例2では、読み順評価手段を設け、該手段はグラフ化の結果から求めた読み順の線分が交わるとき、偽と判定し、読み順再設定手段は、読み順の再設定を行う。再設定は、各本文領域の外接矩形がより左上にある順にソートする。これにより複雑な配置の文書においても正しい読み順が得られる。さらに、実施例3では、実施例1の方法によって抽出された読み順に従って文字認識を行って文書情報を得る。

【0024】

【実施例】以下、本発明の一実施例について図面を用いて説明する。

〈実施例1〉図1は、本発明の実施例1のブロック構成図を示す。図1において、画像入力手段101は文書を2値画像として入力するためのスキャナ等である。領域抽出手段102は、入力画像から文章領域、図領域、罫線領域等の領域を抽出する手段である。文章領域分別手段103は、領域抽出手段102により抽出された文章（文字）領域を、読み順設定の対象である本文領域と、それ以外の領域（図題、表題、ヘッダ、フッタ等）に分別する手段である。罫線情報生成手段104は、領域抽出手段102により抽出された罫線領域や、白領域の架空罫線、図領域の端部の架空罫線などを生成する手段である。文章領域配置構造化手段105は、本文相当領域（囲み枠内も含む）を木グラフとして構造化する手段である。読み順抽出手段106は、木グラフから文章領域の読み順を抽出する手段である。108は以上の各手段を制御する制御部、107は入力画像や抽出した領域、作成した構造の情報等の各種データを記憶するためのデータ記憶部である。109はデータ通信路である。

【0025】なお、102乃至106の各手段は、それぞれ個別のハードウエア手段または個別のソフトウエア手段として実現されてもよいし、共通のハードウエア上でソフトウエアにより実現されてもよい。

【0026】以下、実施例1の動作及び処理内容について、図2の処理フローチャートに従って、図3乃至図1

0を適宜参照しつつ説明する。

【0027】処理ステップ201；画像入力手段101によって、処理すべき文書を2値画像として入力する。この入力画像のデータはデータ記憶部107に記憶される。

【0028】処理ステップ202；領域抽出手段102によって、入力画像の文章（文字）領域、図領域等を抽出する。抽出された領域の情報はデータ記憶部107に記憶される。

【0029】処理ステップ203；文章領域分別手段103による処理ステップであり、抽出された文章領域を図題、表題、ヘッダ、フッタの領域と、それ以外の領域である本文領域とに分類する。この本文領域とは、読み順の設定されるべき文章領域である（枠で囲まれた領域も本文領域として扱うが、枠で囲まれていない本文領域より読み順が後に来るものとする）。以下、この分類の処理について詳細に説明する。

【0030】まず図題、表題を分類する。この処理において、処理ステップ202で抽出された領域の表現が外接矩形のみであった場合には、各領域の図や絵等の実体（イメージそのもの）と、その外接矩形との相違が大きいことがある。このような場合には、図領域と他の文章（文字）領域との重なり等により図の外接矩形を分解して、いくつかの外接矩形の集合によって図の実際に存在する範囲と、その外接矩形による表現との相違を少なくする。

【0031】図題、表題は、図・表の近傍に存在する行数の少ない文章領域である。そこで図と文章領域との距離を計算する。図の輪郭形状が判明しているならば、その図と文章領域との距離を計算し、図が外接矩形で表現されているときは、その外接矩形と文章領域との距離を計算する。そして、この距離が小さく、かつ行数の少ない文章領域を図・表題の候補とする。

【0032】次に、図・表題の候補で、当該文章領域にとって図・表領域の反対側に存在する文章領域との位置関係を調べる。これを図3によって説明する。図3において、301は図領域、302は図領域301の近傍にある図題候補、303は図題候補の文章領域である。この例のように、図題候補302の反対側に文章領域303があり、両領域302，303の左右位置が揃っている場合には、両領域302，303を連続した本文領域と判断し、図題候補302を図題とは分類しない。左右いずれかでも位置が揃っていないときには図題候補302を図題に分類する。ただし、左右の一方の位置が揃っている場合にも図題としない、という方法も採用可能である。

【0033】以上の図題、表題の分類処理に続いて、ヘッダの分類を行なう。ここで、本文領域が縦書きであるか横書きであるかが判定している場合には、そのいずれであってもヘッダは原稿の上部に存在する。縦書き原稿

でも、ヘッダは横書きで原稿上部に存在するのが普通である。また、行（文字列）方向が判明していても、文字の方向が縦か横か分かっていない場合については、行方向が横であっても原稿は縦書きであることがある。原稿を90°回転して入力した時に、そうなる。このような場合でも、画像の上の方が文章の先頭にくるものとすると、画像左側が原稿の上部にあたる。

【0034】このような考察に基づき、行方向が横の場合には画像の上部及び左部についてヘッダの存在を調べ、行方向が縦の場合には画像の上部及び右部についてヘッダの存在を調べる。

【0035】より具体的に述べる。調べる部位に対して、まず罫線の存在を調べる。罫線が存在する場合、この罫線の長さが画像の幅または高さに対して十分に大きく、かつ、この罫線より外側に大きな文章領域（数行を含む文章領域）が存在しないならば、この罫線を本文とヘッダ部とを分ける罫線であると判断する。そして、その外側に小さな文章領域があれば、それをヘッダとして分類する。

【0036】図4に示す例で説明すると、401は入力画像、402は罫線、403～405は文章領域である。行方向が横であることのみ判明しているとすれば、ヘッダは上部または左部に存在する筈であるから、この位置で十分に長い罫線を探す。図4の例においては、罫線402が存在するので、その上側に大きな文章領域が存在するか調べる。文章領域403は数行を含むような大きな領域ではないので、罫線402は本文とヘッダを分ける罫線であるあると判断する。したがって、この罫線402より上側にある小さな文章領域403をヘッダとして分類することになる。

【0037】該当する罫線が存在しない場合、文章領域の存在する範囲の最上部及び最左部から、ある距離だけ内側にはいった位置に架空の罫線を生成し、同様の方法でヘッダの分類を行なう。

【0038】図5に示す例で説明すると、501は入力画像、502は文章領域の存在範囲、503～506は文章領域である。この例では、画像の上部と左部に架空の罫線507，508を生成することになる。この例では、上部の架空罫線507の上側には大きな文章領域は存在しないので、この架空罫線507は、本文とヘッダとを分ける罫線として有効である。そして、この架空罫線507の上側にかさな文章領域503があるので、これをヘッダとして分類する。架空罫線508の左側には大きな文章領域は存在しないが、大きな文章領域504と架空罫線508が重なっている。

【0039】ヘッダの抽出率を上げたい場合には、この架空罫線508のような文章領域と重なった罫線も有効な罫線として扱ってよい。しかし、架空罫線と重なった文章領域がヘッダとして誤抽出されるのを防ぎたい場合には、そのような罫線を無効とすればよい。なお、この

例では画像の傾き（スキュー）がないが、傾きがある場合には、その傾き角度にあわせて罫線を傾けて生成する。

【0040】行方向が縦の場合も同様に、ヘッダと本文を分ける罫線を探索し、罫線がないときの架空罫線を生成して、ヘッダの分類判定を行なう。

【0041】以上のヘッダの分類と同様にしてフッタの分類を行なう。文字方向が判明している場合には、原稿の下部に相当する位置についてフッタを調べる。行方向のみ判明している場合には、行方向が横であれば画像の下部と右部について調べ、行方向が縦であれば画像の下部と左部について調べる。

【0042】以上のようにして分類されたヘッダ、フッタ、図・表題を除く領域が本文領域となる。ただし、囲み枠が存在する場合には、枠内の文章領域を、その枠毎に分類し、枠外の本文領域とは区別しておく。

【0043】処理ステップ２０４；罫線情報生成手段１０４により架空罫線の生成を行なう処理ステップである。ここでいう架空罫線とは、処理ステップ２０３におけるヘッダ、フッタの分類のための架空罫線を除くもので、文章領域の配置構造を表わす木グラフを構築するために図や白領域から新たに生成されるものである。

【0044】まず、図、表等の領域について説明する。なお、行方向を横として座標系をとったとして以下の説明を行なう。ここでは、図及び表領域の左右の端に垂直架空罫線を生成する。図の存在範囲を外接矩形で表現している場合には、図題の分類時に外接矩形の分割を行なっているので、この分割された図領域について架空罫線の生成を行なう。

【0045】図６を例に説明すれば、６０１は表領域、６０２は図領域（の外接矩形）、６０５は図領域６０２と重なった文章領域、６０３と６０４は図領域６０２を分解した領域である。この例では、領域６０１、６０３、６０４の左右端に架空罫線６０６～６１１をそれぞれ生成することになる。

【0046】次に白領域から生成する架空罫線について説明する。この罫線は文字列（行）方向のものであり、ここでは行方向を横としているので水平罫線となる。この罫線の生成（抽出）は、縦横への射影をとる方法によって、あるいは、画像の行方向へのランレングス符号化をして、ある閾値以上の長さを持つ白ランの連結成分を抽出し、この白連結成分の中から水平罫線を十分に構成し得るものを、その幅と高さによって選択し、選んだ白連結成分の中心付近に水平架空罫線を生成する方法によって行なうことができる。

【0047】また、座標系の一番上部に、画像の幅（行方向が縦の場合は座標系を９０°回転しているので画像の高さ）に等しい長さを持つ水平架空罫線を生成する。

【0048】なお、囲み枠線の４辺の線分のうち、上部の線分は水平罫線として扱い、左右の線分は垂直罫線として扱う。

【0049】処理ステップ２０５；文章領域配置構造化手段１０５により、囲み枠外の本文領域の配置構造グラフを作成する処理ステップである。配置構造は木グラフで表わされるので、あるノードが、どのノードの子に相当するかを順次決定していくことによって木グラフを作成することになる。

【0050】まず、ノードとして、囲み枠外の本文領域、水平罫線（架空罫線を含む）を登録する。そして、このノードを上部にあるものから順次処理する。

【0051】今、あるノードに着目しているとすると、この着目ノードより処理順番が後になるノードは着目ノードの子候補となる。ここで、子候補が子ノードに相当するかどうかの判別処理を行ない、子ノードに相当する場合は着目ノードとの間に親子のリンクを張る。ただし、この子に相当すると判別されたノードが既に他のノードの子ノードとしてリンクされていた場合には、どちらが親ノードとしてふさわしいか判別処理を行ない、ふさわしいと判断された方の親ノードと親子関係のリンクを張り、どちらとも判別がつかない場合には木グラフのルートに直接つなぐようにする。また、着目ノードは、それより前に処理したノードの子候補となっている等で、前のノードの処理が完了しているにも拘らず、どのノードの子としても未だリンクされていない場合には、着目ノードを木グラフの子ノードとする。ただし、囲み枠の上部の水平罫線は直接に木グラフのルートの子ノードとする。

【0052】図７の例によって、より具体的に処理を説明する。図７において、７０１は最上部に生成された架空罫線、７０２～７０５は文章領域、７０６～７１０は各ノードの勢力範囲（後述）、７１１と７１２はそれぞれ架空罫線７０１と文章領域７０２の子ノードの探索範囲、７１３は図領域、７１４と７１５は架空垂直罫線、７１６は文章領域７０４の一時的な勢力範囲、７１７は文章領域である。なお、以下の説明において、領域を示す符号を、それに対応するノードを示すためにも便宜用いる。

【0053】まず、最上部のノード（７０１）が最初の処理ノードとなる。このノードは親ノードが未定であるので、木グラフのルートの子ノードとする。ここで、各ノードは勢力範囲と探索範囲を持つ。勢力範囲は親から継承するもので、探索範囲は勢力範囲と最初は等しいが、順次更新されて狭まっていく。

【0054】さて、最初の処理ノード（７０１）は、親がルートであるので、それ自体の幅に等しい勢力範囲７０６を持つとする。そして、この勢力範囲と等しい幅を探索範囲として以下のノードの探索を行なう。

【0055】ノード（７０１）の探索範囲７１１内にノード（７０２）が見つかるので、このノード（７０２）はノード（７０１）の子ノード候補となる。そこで、ノ

9

ード（７０１）の探索範囲７１１のノード（７０２）の範囲（７１１の黒部分）を探索済みとして、以下の探索の範囲から除く。ノード（７０２）はノード（７０１）の子であるので、ノード（７０１）と同じ幅の勢力範囲７０７を継承する。

【００５６】ノード（７０１）の残りの探索範囲で探索すると、ノード（７０４）が見つかる。しかし、ノード（７０４）はノード（７０２）との間でも親子関係がなりたつので、ノード（７０１）との間で親子関係のリンクは張らない。また、ノード（７０３）もノード（７０１）の探索範囲下にあるが、ノード（７０４）と同様にノード（７０２）の勢力範囲下にありノード（７０２）と親子関係がなりたつので、ノード（７０１）とノード（７０３）は親子ではない。ノード（７０５）は僅かながらノード（７０１）の勢力範囲下にあるが、探索範囲は連続したある程度の幅のみ有効とするので、探索外となる。

【００５７】次にノード（７０２）が処理ノードとなる。まず、ノード７０４がノード（７０２）の子ノードとなる。ここで、勢力範囲は垂直罫線を越えないものとする。したがって、ノード（７０４）の勢力範囲は垂直罫線７１５を越えない７１６の範囲となる。次にノード（７０５）がノード（７０４）の勢力範囲下にあり、探索範囲７１２にも含まれるが、ノード（７０４）とノード（７０５）は親子的位置関係にないので、ノード（７０５）はノード（７０２）の子ノードとなる。複数の子ノードがある場合には、その勢力範囲を適当な位置で分割する。ここでは勢力範囲を中点で分割するものとすると、ノード（７０４）とノード（７０５）の勢力範囲は、７０９の範囲と７１０の範囲に分割される。また、次にノード（７０３）もノード（７０２）の探索範囲下にあり、子ノードとなって勢力範囲を継承する。

【００５８】なお、この例ではみられないが、親ノードの勢力範囲を子ノードの領域が越える場合には、子ノードの勢力範囲を継承した範囲から、その越えた分だけ拡張する。また、囲み枠の上部の線分を水平罫線としてノードにしているが、このノードの勢力範囲は親から継承するのではなく、それ自体の幅に等しい範囲とする。次にノード（７０４）が処理ノードとなる。ノード（７０４）の探索範囲７０９内にノード（７１７）が存在する。したがって、ノード（７１７）はノード（７０４）の子ノードとなり、探索範囲は全て満たされるので、次にノード（７０５）のの処理に移る。

【００５９】ノード（７０５）の探索範囲７１０内にもやはりノード（７１７）があるが、このノード（７１７）は既にノード（７０４）の子ノードとなっているので、ここで親ノードの選択を行なうことになる。ところが、ノード（７０４）とノード（７０５）は同じような幅を持ち、ノード（７１７）は両方の勢力範囲に十分にはいっているため、どちらが親であるが一意に定まらない。そこで、ノード（７１７）をノード（７０４）の子から外し、あらためてルートの子ノードとする。なお、子ノードが複数ある場合、木グラフでの左右の並びは、例えばノード（７０３）、ノード（７０４）、ノード（７０５）のように普通のノードに接続する場合は、その位置通りにノード（７０３）を一番左側、ノード（７０５）を一番右側にする。ノードの子ノードの場合は、新しく子ノードを接続する度にとりあえず一番右側へと接続しておき、最終的にソートする。

10

【００６０】続いてタイトル部位の処理について、図７及び図８に示した例によって説明する。タイトル部位の処理は、図７に関連して説明した処理の中で行なわれる。各ノードには、タイトル部位であるか否かを示すタイトルフラグを付ける。図７の例でいえば最初のノード（７０１）のタイトルフラグは必ず立て（オンし）、タイトル部位であるとする。次にノード（７０２）がノード（７０１）の子ノードとなるわけであるが、この時に、ノード（７０２）の左右に文章領域が存在するか調べる。左右に文章領域が存在しなければ、ノード（７０２）のタイトルフラグも立てる。ただし、後にノード（７０１）の他の子ノードがリンクされた場合には、ノード（７０２）はノード（７０１）の唯一の子ノードでなくなるので、ノード（７０２）のタイトルフラグを下ろす（オフする）。

【００６１】ノードのタイトルフラグが立っている場合には、左側の勢力範囲を架空垂直罫線で抑えられているときに、その架空罫線を１回だけ無視する形で、それを越えて拡大する。図８の例で説明する。図８において、８０１は最上部の架空水平罫線、８０２は文章領域、８０３は図領域、８０４と８０５は図領域８０３の両端に生成された架空垂直罫線、８０６は架空水平罫線８０１の勢力範囲、８０７は文章領域８０２の勢力範囲である。この勢力範囲８０７は、架空垂直罫線８０５に遮られず、次の架空垂直罫線８０４まで延ばされる。ノード７０２の勢力範囲７０７は、もともと遮るような垂直罫線が存在しないので関係はない。

【００６２】さて、ノード（７０２）の子ノードの探索に移ると、まずノード（７０４）が探索されることは前述のとおりであるが、ノード（７０４）の左右には文章領域７０３，７０５が存在する。したがって、これ以降の文章領域はタイトル部位とはならないので、ノード（７０２）のタイトルフラグを下ろす。親のノードのタイトルフラグが立っていない場合には、その子ノードがタイトル部位であるか否かを調べるために、左右に文章領域が存在するか探索する必要もなくなり、以下の処理時間が短縮される。

【００６３】以上の処理を　ノードに相当する文章領域（囲み枠外のもの）について、上から下まで全てについて行なう。

【００６４】処理ステップ２０６；文章領域配置構造化

手段１０５により、囲み枠内の文章領域をグラフ化する処理ステップである。

【００６５】基本的には先の囲み枠外の本文領域のグラフ化処理と同様に、処理ノードの勢力範囲下にあるノードを、処理ノードの子ノードとしてリンクしていく。先の囲み枠外本文領域のグラフ化処理でリンクが張られたノードは木グラフを構成している。この中には囲み枠線の上部の水平線分もノードとして登録されている。したがって、そのノード毎に、その囲み枠内の文章領域を対象に木グラフを構成する。

【００６６】図９の例で説明する。図９において、９１０～９１６は文章領域、９１７と９１８は囲み枠である。９０１はルート、９０２～９０７は当該処理ステップ２０６の前に登録されたノードである。９０２は最上部架空罫線に相当するノード、９０３は文章領域９１０に相当するノード、９０４は文章領域９１１に相当するノード、９０５は文章領域９１２に相当するノード、９０６は囲み枠９１８の上部水平罫線９２０に相当するノード、９０７は囲み枠９１７の上部水平罫線９１９に相当するノードである。

【００６７】囲み枠の上部水平罫線９１９，９２０は、先の処理ステップ２０５における最上部罫線（９０２）に対応し、それと同様の処理を行なうことになる。ただし、水平罫線９１９，９２０のタイトルフラグは常にオフにしておく。したがって、囲み枠内では、タイトル部位の処理は行なわれない。

【００６８】図９の例では、ノード９０６（つまり水平罫線９２０）の下には文章領域９１３があるので、文章領域９１３を子ノードとしてノード９０６に接続する。また、ノード９０７（つまり水平罫線９１９）の下に文章領域９１４があるので、これを子ノードとしてノード９０７に接続し、また文章領域９１４の下には文章領域９１５，９１６があるので、これら二つの領域も文章領域９１４に子として接続する。ここでも、先の処理ステップ２０５での処理と同様に勢力範囲及び探索範囲を用いて子ノードの探索、及び子ノードであるか否かの判別を行なう。ただし、ルート９０１に直接接続されたノード９０６，９０７の勢力範囲は、それ自体の幅に等しい。

【００６９】以上のようにして木グラフを作成したならば、次にルートの子ノードのソートを行なう。ルートの子ノードになっているのは、最上部の架空水平罫線、親ノードが一意に定まらなかった領域、あるいは囲み枠の上部罫線である。このようなノードのうち、囲み枠の上部罫線は、他のノードよりも木グラフ上で右側に来るようにソートする。また、囲み枠のノード同士、及び、囲み枠の罫線同士については、より上に位置するもの、より左側にあるものを、グラフ上でより左側にするように順番を入れ替える。この際、各ルートの子ノードの勢力範囲を使用することによって、どちらが上位にあるかを判別することができる。

【００７０】処理ステップ２０７；文章領域配置構造化手段１０５により、図・表題の分別を再度行なう。ここでは、グラフの葉にあたるノード（子を持たないノード）が罫線ではなく、文章領域であって、その行数が少なく（この行数の閾値は、先の処理ステップ２０３における図・表題の分別に使用したものと同様の値でよい）、かつ、その親ノードの実体との間に、ある程度大きな図が存在する場合には、当該文章領域を図題または表題と分別し、これを本文配置の木グラフから取り除く。

【００７１】図１０の例によって説明する。図１０において、１００１と１００２はノード、１００３と１００４はそれぞれノード１００１，１００２の実体である文字領域である。１００５は図領域である。この例のノード１００２は葉に相当するもので、その実体たる文章領域１００４の行数が少ない。また、その親ノードたるノード１００１との間に、比較的大きな図領域１００５が存在する。したがって、ノード１００２は本文配置の木グラフから取り除かれる。

【００７２】処理ステップ２０８；読み順抽出手段１０６において、以上の処理で得られた本文配置を示す木グラフ上で先行順探索を行ない、罫線やルートを除いた文章領域の順番を、本文領域の読み順として抽出する。

【００７３】実施例１の一態様を示すと以下のようになる。配置構造を表す木グラフを構築する際に、文章領域を木グラフのノードに割り当て、各ノード毎に他のノードへの配置関係を表わす勢力範囲を求め、各ノードの親子関係の探索を勢力範囲に従って行ない、子ノードに親ノードの勢力範囲を継承させることによって勢力範囲の更新を行ない、親子ノードの探索を繰り返すことによって木グラフを構築する。このように、勢力範囲を用いて親子ノード探索を制御することにより、タイトル部位や図等に適切に対処して本文領域を適切に構造化し、本文領域の読み順を正しく設定できる。

【００７４】文章領域以外の特定の領域も木グラフのノードに割り当てる。例えば、文字列方向と同方向の罫線を抽出し、これを木グラフのノードに割り当てる。また、白画素の領域で、ある値より文字列方向に長い白画素領域を抽出し、これを木グラフのノードに割り当てる。このように、文書上の文章領域以外の罫線等の様々な要素、例えば文字列方向と同方向または垂直方向の罫線や、空白部分、図・表領域等による文章領域の配置への影響を適切に扱うことにより、そのような要素が存在する文書の本文領域の読み順を正しく設定できる。

【００７５】文字列方向と垂直の方向の罫線によって、ノードの勢力範囲を制限する。また、図・表領域の文字列方向についての両端に文字列方向と垂直の架空罫線を生成し、この架空罫線によってノードの勢力範囲を制限し、架空罫線による勢力範囲の制限をタイトル部位にお

いて変更する。このように、図・表領域の文字列方向の両端に垂直の架空罫線を生成し、この架空罫線によってノードの制御範囲を制限することにより、文字列方向と垂直方向の罫線による配置への影響を適切に処理することが可能になり、また、図・表領域による本文領域の配置への影響を適切に処理して、本文領域の読み順を正しく設定できる。タイトル部位において架空罫線による勢力範囲の制限を変更することによって、タイトル部位の配置を適切に処理することができる。

【００７６】文章領域の配置構造を木グラフとして構築し、構築された木グラフと文章領域以外の図等の領域の位置情報とに基づいて誤って本文領域と分類された領域を判別し、それを木グラフから取り除き、この処理の後の木グラフに従って文章領域の読み順を設定する。このように、本文領域の木グラフを作成した後に、本文領域とそれ以外の領域との細分別を行なうことによって、本文と図・表題を高精度に分別し、より正確な文章領域の構造化と読み順設定が可能となる。

【００７７】〈実施例２〉上記した実施例１は、構造化に失敗したり、もともと木グラフで表現することに適さない構造をもつ文書である場合に、構造化できない部分のみならず、全体的に読み順が変更されてしまう可能性がある。そこで、本実施例２では、文章領域の配置構造を求めて読み順を求めた後、該読み順の評価を行うようにした。

【００７８】図１１は、本発明の実施例２のブロック構成図である。図において、１１０１は、画像の入力手段、１１０２は、入力画像から領域を抽出する領域抽出手段、１１０３は、抽出された文字領域を、本文領域とそれ以外の図題、表題、ヘッダ、フッタ等に分別する文章領域分別手段、１１０４は、本文相当領域を木グラフとして構造化する文章領域配置構造化手段、１１０５は、木グラフから読み順を抽出する読み順抽出手段、１１０６は、読み順抽出手段１１０５で抽出した読み順を評価する読み順評価手段、１１０７は、評価結果が偽であるときの読み順を再設定する読み順再設定手段である。１１０８は、入力画像や抽出した領域、作成した構造の情報などの各種データを記憶するデータ記憶部、１１０９は、上記した各手段を制御する制御部、１１１０は、データ通信路である。

【００７９】図１２は、実施例２の処理フローチャートである。以下、実施例２の動作を説明すると、まず、スキャナなどの画像入力手段１１０１によって文書を２値画像として入力する（ステップ１２０１）。次いで、この２値画像から文字領域、図領域などの領域を抽出する（ステップ１２０２）。この抽出方法としては、例えば特開平５－８１４７５号公報に記載された文字領域抽出方法などを用いればよい。

【００８０】文章領域分別手段１１０３は、抽出した領域を、図題、表題、ヘッダ、フッタと、それ以外の本文

領域に分類する（ステップ１２０３）。ここで、本文領域とは、読み順の設定される領域で、後述するように枠で囲まれた領域についても、枠で囲まれていない領域より読み順が後にくる本文領域として処理される。

【００８１】文章領域配置構造化手段１１０４は、本文領域の配置を構造化し、木グラフで表現する（ステップ１２０４）。そして、読み順抽出手段１１０５は、このグラフ表現から先行順探索で読み順を決定する（ステップ１２０５）。なお、ステップ１２０３～１２０５の処理については、前述した実施例１に記載の方法を用いる。

【００８２】その後、読み順評価手段１１０６は、読み順の評価を行う（ステップ１２０６）。図１３は、本実施例に係る読み順の評価を説明する図である。図において、１３０１から１３０７は、抽出された本文領域の外接矩形である。まず、この各外接矩形の中心点を求める。１３０８から１３１４は、求められた中心点である。そして、この点を読み順に従って線分で結ぶ。

【００８３】いま、グラフ化の結果から求めた読み順が図１３に示すものであったとすると、中心点１３１２と中心点１３１３との間の線分１３１２－１３１３と、中心点１３１０と中心点１３１４との間の線分１３１０－１３１４とが交わるため、この結果の評価は偽と判定される。

【００８４】従って、読み順再設定手段１１０７は、読み順の再設定を行う（ステップ１２０７）。再設定は、各外接矩形の位置に着目し、より左上にある順にソートする。例えば、外接矩形１３０１の下部は外接矩形１３０２や１３０５の上部よりも上にある。このような場合、外接矩形１３０１は外接矩形１３０２、１３０５よりも"上"と判断する（外接矩形１３０１>外接矩形１３０２、外接矩形１３０１>外接矩形１３０５）。

【００８５】また、外接矩形１３０２の右部は外接矩形１３０５の左部よりも左にあるので、"左"と判断する（外接矩形１３０２>外接矩形１３０５）。なお、このときの判断には余裕を持たせるようにしてもよい。外接矩形１３０１と１３０２の左右の位置を比べた場合には どちらが左とも右とも判定できない。そこで、３つの外接矩形１３０１、１３０２、１３０５を比較すると、上記した関係から外接矩形１３０１>外接矩形１３０２>外接矩形１３０５となる。

【００８６】このような判定方法を全ての領域に適用することにより読み順を決定する。外接矩形１３０５と１３０３を比べた場合には、上下では外接矩形１３０５>１３０３であり、左右では外接矩形１３０３>１３０５となる。このような場合には左右の関係を優先する。従って、外接矩形１３０３>外接矩形１３０５となる。

【００８７】上記した判定処理の結果、最終的には図１４に示すような読み順が得られる。図１４において、１４０１から１４０７は外接矩形、１４０８から１４１４

15　　16

は中心点、線分１４０８－１４０９－．．．－１４１４は読み順を示す。

【００８８】なお、本実施例の読み順の評価および再設定は、囲み枠内のものについては、その枠内の領域に対してのみ行い、枠外および他の枠内のものとは区別して処理する。図１５は、囲み枠線を有する本文領域の例を示す。図において、１５０１から１５０６は本文領域、１５０７は本文領域１５０４と１５０５を囲む囲み枠線、１５０８は本文領域１５０６の囲み枠線、１５０９から１５１４は各領域の中心である。

【００８９】図１５の本文領域に対して、図１２のステップ１２０１からステップ１２０６の処理の結果、線分１５１３－１５１４が他の線分と交わっていると評価される。しかし、この場合、線分１５１３－１５１４は異なる枠内領域１５０７と１５０８を結ぶ線分であることから、評価対象外となり、再設定を行わない。評価対象となる線分は、線分１５０９－１５１０－１５１１、線分１５１２－１５１３の線分である。

【００９０】（実施例３）本実施例は、実施例１の方法によって抽出された読み順に従って本文領域について、例えば文字認識を行って文書情報を得るようにしたものであり、実施例１によって抽出された文書情報の利用形態に係る。

【００９１】図１６は、実施例３のブロック構成図である。画像データは、スキャナなどの入力手段１６０１あるいは、回線に接続されたファクシミリなどの信号受信手段１６０２から入力される。情報抽出処理手段１６０３は、実施例１で説明した文章領域の抽出、文章領域の判別、文章個域の配置構造化、読み順の抽出の他に、更に文字認識を行って、文書情報を得る機能などを備えている。

【００９２】表示手段１６０４は、文章領域を表示する例えばＣＲＴディスプレイであり、修正指示手段１６０５は、表示された抽出情報などに誤りがあった場合に修正する例えばマウスなどのポインティングデバイスであり、結果出力手段１６０６は、紙などに出力するプリンタ、あるいは電子情報として媒体に格納する蕃積手段、通信回線を介して伝送する伝送手段である。

【００９３】図１７は、情報抽出処理手段１６０３の構成を示す。ここで、領域抽出手段１７０２、文章領域分別手段１７０３、罫線情報生成手段１７０４、文章領域配置構造化手段１７０５、読み順抽出手段１７０６、データ蓄積部１７０８、制御部１７０９、データ通信路１７１０は、それぞれ実施例１で説明したものと同一の機能、構成を有している。本実施例では、これら手段に加えて、入出力データが格納され、バッファとして機能するデータ入出力手段１７０１と、文字認識手段１７０７と、文字以外の領域処理手段１７１１が設けられている。

【００９４】実施例１で説明したように、領域抽出手段１７０２は入力画像から文章領域とそれ以外の表領域、図の領域などを抽出し、文章領域分別手段１７０３は本文とそれ以外とを分別し、本文領域について、罫線情報生成手段１７０４、文章領域配置構造化手段１７０５、読み順抽出手段１７０６は、その配置構造と読み順を抽出する。

【００９５】文字認識手段１７０７は、上記した本文領域の読み順に従って文字認識を行って、文書情報を出力する。また、本文以外の文字領域についても文字認識を行って、文書情報を出力する。文字以外の領域処理手段１７１１は、表などの文章領域以外について、適応処理を行う。具体的には、表であれば罫線情報を抽出し、表内文字を文字認識手段１７０７で文字認識する。写真であれば、例えば２次元ＤＣＴ変換によって適応符号化を行い、図であれば、例えば線画を抽出してベクトル化を行う。これら文章以外の領域についての処理は、文章領域の処理とは別に独自に行ってもよい。このように、抽出した文章領域以外については領域の特性に合った処理をしているので、表や写真、図などの情報を最適な形で利用することができる。また、実施例１で説明したように、本文以外のヘッダ、フッタなどの情報を抽出することにより、本文以外の文章情報を書誌的情報として取り出して利用することができる。

【００９６】上記したようにして抽出された文字情報は、表示手段１６０４に表示される。図１８は、文章領域の表示例を示し、文章領域は矩形で表示され、各領域には読み順に従った番号が同時に表示される。すなわち、図１８において、表示された文書画像１８０１には、ヘッダ領域１８０２、本文領域１８０３～１８０５、図の領域１８０６が表示される。そして、各本文領域１８０３～１８０５には読み順１８０７（番号は１）、１８０８（番号は３）、１８０９（番号は２）も表示されている。なお、読み順の表示方法としては、この他に各本文領域を読み順通りに矢印で結ぶ表示形式を採ってもよい。

【００９７】表示された抽出情報、読み順などに誤りがあった場合、修正指示手段１６０５を用いて修正する。例えば、マウスなどで領域の大きさ、位置を変更したり、あるいは指定した領域の読み順をキーボードなどから入力などして修正する。抽出した文字情報は修正、確認した後、文書情報が確定し、結果出力手段１６０６に出力される。このように、文書情報抽出処理において誤った処理が行われても、その結果を確認、修正しているので、最終的に文書情報をより迅速に得ることができる。

【００９８】

【発明の効果】以上、説明したように、請求項１記載の発明によれば、文章領域を本文領域とそれ以外の領域とに分別して本文領域の配置構造を求めているので、本文領域とは異なる読み順を持つ文章領域の影響を排除し

17

て、本文領域の読み順を適切に抽出することができる。

【0099】請求項2記載の発明によれば、横書き／縦書きに共通な本文領域の配置構造を求めているので、横書き文書も縦書き文書も同一の処理によって本文領域の正しい読み順を抽出することが可能となる。

【0100】請求項3記載の発明によれば、本文領域を木グラフのタノードに割り当てることによって本文領域の配置構造を求めているので、本文領域の配置の大局的な構造が表現され、これによって段落の有無や段落の形態に依存しない本文領域の読み順を抽出することができる。

【0101】請求項4記載の発明によれば、囲み枠の内部の文章領域を、囲み枠の外部の文章領域から区別し、別個に読み順を抽出しているので、文章領域の囲み枠の存在する文書の場合にも、囲み枠の内部の文章領域も含めて文章領域の読み順を正しく抽出することができる。

【0102】請求項5記載の発明によれば、文章領域の配置構造を求めて読み順を求めた後、該読み順の評価を行い、この評価が偽の判定であった場合には、読み順の再設定を行うようにしているので、複雑な配置の文書であっても、正しい読み順が得られると共に、仮りに、最初に誤った読み順が抽出されたとしても利用者による修正負担を少なくすることが可能となる。

【0103】請求項6記載の発明によれば、読み順の評価方法として、各領域に基準点を設けて、この基準点間を読み順に線分で結んだときに、線分に交わりが生じた場合に判定を偽としているので、配置構造を木グラフで表現した結果が不自然な読み順を与えていることを判別することができる。その判別結果から読み順の再設定を行うことによって、利用者による修正負担を少なくすることが可能となる。

【0104】請求項7記載の発明によれば、囲み枠内の本文領域は、それぞれの枠内の領域について読み順の評価および再設定を行っているので、異なる枠内領域を結ぶ読み順の線分が枠外本文領域間を結ぶ読み順の線分と交差することに無関係に、線分の交差という基準に基づいて、読み順が正しいか否かを判別することが可能となる。

18

【0105】請求項8記載の発明によれば、複雑なレイアウトの文書が入力されても、入力された文書画像から正しい読み順で文字認識を行っているので、正確な文書情報を抽出し、利用することができる。

【図面の簡単な説明】

【図1】本発明の実施例1のブロック構成図である。

【図2】実施例1の処理フローチャートである。

【図3】図・表題の分別の説明図である。

【図4】ヘッダの分別の説明図である。

【図5】ヘッダ分別のための架空罫線の生成の説明図である。

【図6】木グラフ構築のための架空罫線の生成の説明図である。

【図7】囲み枠外の本文領域の構造化の説明図である。

【図8】架空罫線と勢力範囲との関係の説明図である。

【図9】囲み枠外の本文領域の木グラフと囲み枠内の文章領域のグラフ化の説明図である。

【図10】図・表題と本文との再分別の説明図である。

【図11】本発明の実施例2のブロック構成図である。

【図12】実施例2の処理フローチャートである。

【図13】実施例2に係る読み順の評価を説明する図である。

【図14】読み順評価の結果、再設定された読み順を示す。

【図15】囲み枠線を有する本文領域の例を示す。

【図16】実施例3のブロック構成図である。

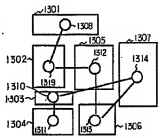
【図17】情報抽出処理手段の構成を示す。

【図18】文章領域の表示例を示す。

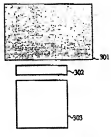
【符号の説明】

101　画像入力手段
102　領域抽出手段
103　文章領域分別手段
104　罫線情報生成手段
105　文章領域配置構造化手段
106　読み順抽出手段
107　データ記憶部
108　制御部
109　データ通信路

【図13】

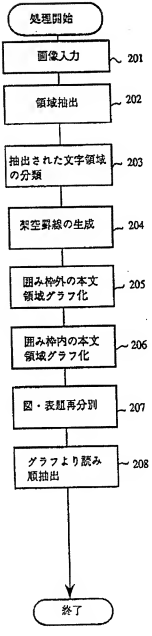


【図2】

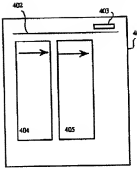


301～303：文章領域

【図3】

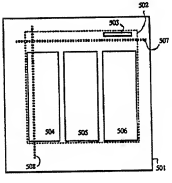


【図4】



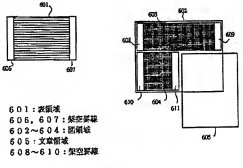
402：罫線
403～405：文章領域

【図5】



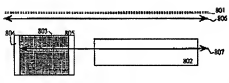
502：文章存在範囲
503～506：文章領域
507，508：架空罫線

【図6】



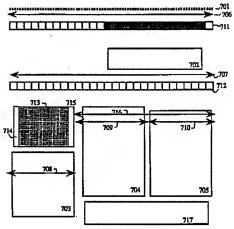
601：表領域
606，607：架空罫線
602～604：図領域
605：文章領域
608～610：架空罫線

【図８】



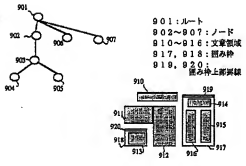
８０１：架空罫線
８０２：文章領域
８０３：表領域
８０４，８０５：架空罫線
８０６：勢力範囲

【図７】



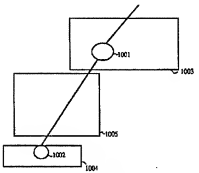
７０１：架空罫線
７０２～７０５，７１７：文章領域
７０６～７１０，７１６：勢力範囲
７１３：表領域
７１４，７１５：架空罫線

【図９】



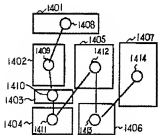
９０１：ルート
９０２～９０７：ノード
９１０～９１６：文章領域
９１７，９１８：囲み枠
９１９，９２０：
囲み枠上部罫線

【図１０】



１００１，１００２：ノード
１００３，１００４：文章領域
１００５：図領域

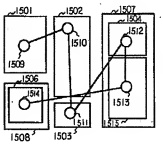
【図１４】



１４０１～１４０７：
本文領域の外接矩形

１４０８～１４１４：
外接矩形の中心点

【図１５】

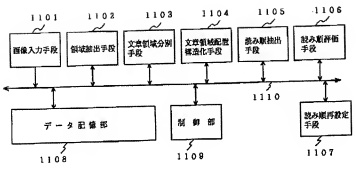


１５０１～１５０９：
本文領域

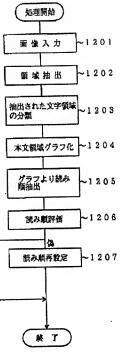
１５０７，１５０８：
囲み枠線

１５０９～１５１４：
中心点

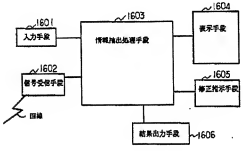
【図１１】



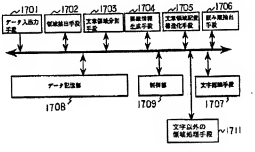
【図１２】



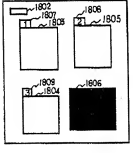
【図１６】



【図１７】



【図１８】



1801：文書画像
1802：ヘッダ領域
1803～1805：本文領域
1806：図領域
1807～1809：誤り候